FEASIBLE MULTI-COMPUTER ARCHITECTURES, GIVEN 3-DIMENSIONAL STACKED WAFERS

by

Leonard Uhr

Computer Sciences Technical Report #518

October 1983

FEASIBLE MULTI-COMPUTER ARCHITECTURES, GIVEN 3-DIMENSIONAL STACKED WAFERS*

Leonard Uhr
Computer Sciences Department, University of Wisconsin

Abstract

Stacks of silicon chips, or of wafers, may soon be technologically feasible (see Etchells, Grinberg and Nudd, 1981, for a description of their new thermomigration techniques for producing feedthroughs and microspring interconnects between the surfaces of adjacent wafers of chips). This paper suggests several potentially very powerful architectures that, given such a technology, will become possible, and attractive. These include stacked SIMD arrays, and also several types of stacked and/or tiled-and-stacked tree, pyramid and network structures, among them systems that combine SIMD with asynchronous MIMD processors.

Introduction: Pin Fanout and Memory Limits, and How They Might Be Overcome

Today's VLSI technologies have already achieved 100,000 and 450,000 transistor-equivalent device chips. Device counts have been doubling every 12 to 18 months for the past 15 years or more. This rate of increase in packing density can be expected to continue for the next 10 to 20 years, since a number of technologies are being developed simultaneously, and they all are in their early stages of maturity, or pre-adolescence. We can therefore look with confidence to chips with 1,000,000 devices, probably well before 1990; and it is likely that we will see chips with 10,000,000 or even 100,000,000 devices by 2001.

Limits on Pins Linking Chips and Devices for Memories

The number of connections to and from each individual chip (for input, output, control, power, etc.) has been severely limited on traditional chips because of the "pin fanout" problem. Today's chip packages can conveniently handle only 40 or 64 pins. Expectations are that, to give economically viable yields, this type of packaging will be limited to 128 or 256 pins at most.

A typical traditional single-CPU serial computer uses very large amounts of memory, containing from 1,000,000 to 100,000,000 bits. Assuming that only 1 or 2 devices are needed to store each bit (as is the case in today's 64,000 bit dynamic RAM memory chips), even with the densest future packings envisioned virtually the entire chip area might well be committed to memory in such a traditional design, and there would still be a need for a number of additional chips dedicated exclusively to memory. For the next 5 to 10 years one computer per chip, along the design lines suggested by Patterson, Fehr and Sequin (1979) for 1986 technology, appears the most plausible to many people. Indeed, even with the 1,200,000 device chip they specified they suggest using several additional RAMs for main memory.

A processor needs at least N pins, or, better, 2N pins, to handle parallel input and output of its N-bit words, plus a

number of other pins to handle control, synchronization and power. So given traditional 32-bit computers only two or four could be put on a single chip (with the consequently reduced amount of memory further reduced because it must be divided between them) and still have reasonably fast access to external memories.

Traditional computers have typically used roughly 10,000 gates for relatively powerful processors (see, e.g., Kuck, 1978), although today many more gates are often used with VLSI technologies for functionally equivalent processors, in order to minimize chip area used and processor complexity rather than device count. So one can conceive of putting 100, or even 1,000 or 10,000 relatively powerful processors on a single chip. (This makes the rather conjectural assumption that processor design will gradually approach the sophistication and packing densities of the much simpler, highly modular memory chips. It also means that radically different designs would be needed, allocating far smaller memory banks to each processor and/or having several or many processors share memories.)

The 1-bit processors in today's large arrays (e.g., CLIP4, Duff, 1978; DAP, Reddaway, 1978; MPP, Batcher, 1980) use only from 100 to 600 gates. But it is important to emphasise that they are indeed still programmable, general purpose computers. They have been stripped clean of special-purpose hardware (e.g., for floating point arithmetic, 8-bit symbol matching), and they must do arithmetic bit-serially. But when we examine their simpler, more modular design, which makes them very attractive candidates for VLSI technologies, and consider the tradeoffs between, e.g., 100,000 100-gate processors and 1,000 10,000 gate processors, it is not at all clear that one or the other is "better." Rather, it seems important to explore which is preferable for a variety of different types of problems and algorithms.

But the pin fanout problem appears to severely constrain the number of either 1-bit or 32-bit computers that might be placed on each chip. Where each 1-bit computer needs only one pin the 32-bit computer needs 32 pins. But 100 simple 1-bit computers, needing 100 pins, can be built from the same number of gates as one traditional 32-bit computer.

Possible Partial Solutions

This quandary will change markedly when either or both of the following two possibilities is realized:

A) Enough processors can be put on a single chip, along with their needed memory, so that they will work long enough without needing input-output to and from the chip for each processor. That is, the chip will largely be a self-contained multi-computer system.

B) Much larger numbers of connections can be established between chips than is possible with today's pin technology.

The continuing increase in chip packing density would appear

to ensure that enough processors, but only of the simple 1-bit type, can be placed on a single chip to do significant computing, at least in several very important application areas. For example, a 64 by 64 array of 4,000 processors (the size of the DAP) would be large and powerful enough to process 512 by 512 Television images, by having each processor serially iterate over an 8 by 8 sub-array of the larger picture. We should be able to achieve such chips within 10 or 15 years.

3-Dimensional Stacks of Wafers

The new technology being developed by Etchells, Grinberg and Nudd, which will allow thousands of connections in the third dimension, between adjacent stacked wafers, each containing many chips, would overcome the second problem. It would therefore make feasible a wide variety of new network architectures for large 32-bit as well as simple 1-bit computers.

Etchells, Grinberg and Nudd intend to build a prototype 32 by 32 system with 10 processing planes. They are using relatively conservative 3 to 5 micron CMOS design rules, running at a 10 MHz clock rate (therefore appreciably larger and faster systems could already be built today). Rather than put the entire 1-bit general-purpose computer on one chip, they use six different wafer types, one for each of six different cell types:

1) memory (store, shift, invert, logical-or, readout),
2) accumulator (store, add, readout),
3) shorted plane (I-O, stack/control, communication in 1 cycle),
4) I-O (serial 100 MHz digital I-O),
5) 1-bit word accumulator (count in/shift out),
6) comparator (store, reference, greater/equal/lower).

They suggest that the first two cell types were chosen to perform the functions needed for the range of algorithms they want to handle, while the last four speed up operations.

The operations they have explored include simple object classification, edge detection, histogramming, fourier transforms, cross correlation, and matrix inversion.

Interesting Possible Architectures for Stacked Multi-Computer Wafers

The circuitry for the memory cell and the accumulator cell (which together effect the bulk of the processing) each appears to use roughly the numbers of gates used by processors in systems like DAP, MPP and CLIP. Each has a 16-bit memory register and enough gates to perform its set of logic operations. Each memory cell connects to the 4 nearest-neighbor cells, whereas accumulator cells connect only to adjacent layers.

Therefore it appears that these combine to form special-purpose processors, but that equivalent numbers of general-purpose processors might be designed to replace them.

Stacked 3-Dimensional Lattices of Special-Purpose Processors

Etchells, Grinberg and Nudd suggest that future 3-D systems should have the same general structure, using their six basic wafer types, stacked in differing orderings and numbers, into what I will call a "lattice." Thus more wafers can be stacked, and many more processors can be put on each wafer with denser future technologies. Their present design uses a single external controller, so that the whole system is an SIMD machine.

The first prototype, with only 10 processing planes, each a very simple 16-bit memory or accumulator slice, can either be thought of as a 2-dimensional array of special-purpose processors or a 2-dimensional pipe-line of very simple systolic-like (see Kung, 1980) processors. But stacks with many more planes might be used to construct 3-dimensional arrays of more or less special-purpose processors, through which 2-dimensional arrays of data could be pumped.

Since the accumulator and memory layers do virtually all of the processing, deep stacks of alternating accumulators and memories would seem an attractive structure through which to pump information. Alternately, one might consider combining accumulator and memory into a single cell on a single wafer, or developing a variety of other special-purpose modules, one for each operation, and stacking a special-purpose sequence appropriate for each particular application.

Stacked 3-Dimensional Lattices of General-Purpose Computers

A relatively straightforward variant would replace the special-purpose modules in the 3-D computer with simple general-purpose 1-bit computers. This might be effected in any of several different ways:

A) Each wafer might contain an array of 1-bit computers, much as is already the case for the CLIP4 chip (8 computers), the DAP chip (4 computers) and the MPP chip (8 computers). Now interconnections in the third dimension between wafers would link whole computers.

B) Alternately, a computer might be modularized, its separate parts realized on separate wafers, as in the 3-D computer.

C) Two types of wafer, one with an array of general-purpose processors, the next with an array of each relatively large dual-port memories might be interspersed, so that each memory was linked to one processor on each of that wafer's adjacent wafers. This would allow processors to pass data and messages via the memories they shared, and free processors from the burden of having to stop everything and immediately accept and process such information. Thus memories would serve as I-O buffers and to handle information flow among computers, as well as performing their traditional functions.

D) As packing densities increase, 32-bit (and also 16-bit and 8-bit) computers could be used. Computers of different sizes, powers and specializations could even be intermixed, if deemed desirable.

Pyramids and Diagonally Linked Lattices

The 3-D computer and the proposed lattice are designed to link components to their 2 square neighbors in each of the 3 dimensions. A simple, and potentially very useful, extension would add diagonal links as well. Here several variants are possible:

A) A 2 by 2 array of 4 computers on one wafer might have each computer link to a single parent computer on the next wafer. Now each computer has 4 links from "children" on the wafer "below" it. This gives a structure reminiscent of a quad-tree (Rosenfeld, 1980) or some of the pyramid/cone structures being investigated by Levine, 1978, Tanimoto, 1976, Uhr, 1972, 1978, and others (see also Tanimoto and Klinger, 1980). It has several major problems:

1) The additional links between wafers may well be the limiting factor.

2) There will be only 1/4th as many computers on each subsequent wafer, and only logN wafers in toto for an N by N pyramid.

B) Each computer might link to each of 4 adjacent computers in the next wafer. Now every computer has 4 links to the wafer below it, and also 4 links to the wafer above it. This system now has the same number of computers on each wafer, but at the cost of almost doubling of the number of links between wafers.

C) Additional links and logic gates might be added to each wafer, to collect and combine the information generated by the 4 computers that would link to the same computer in the next wafer in schemes A) and B) above. (This new node might be a special-purpose processor - e.g., to compute the mean, min, or max - or it might also be a general-purpose computer.) Now these new nodes could be linked from wafer to wafer. If only these new nodes were linked, there would be a 4-fold reduction of between-wafer links over the original 3-D computer, yet with both the capability to converge information as in a pyramid and also a full complement of computers on each wafer. This then suggests the possible value of somewhat specialized processors to collect and converge information generated by adjacent computers on the same wafer, in order to then ship this information along to the next wafer.

D) No special new nodes are needed; instead the general-purpose computers could themselves collect and combine the information to be shipped to the next layer. This might be done either by serially polling each of the cells involved (which takes time), or by adding additional hardware (as found in CLIP4) to fetch and combine the needed information in parallel (which takes additional gates, and space). Here are several tradeoffs between adding special accumulators between cells (which specializes the system to a fixed convergence pattern) and using the general-

purpose computers for this purpose (which ties them up and slows them down, and forces us to give all of them additional hardware that only a few will use only some of the time).

Pyramids and Cones of Different Shapes

Rather than 2 by 2 to 1 convergence from children to parent, any of the above systems could be built with 3 by 3 convergence or 7-fold hexagonal convergence (in both cases to a parent that is "centered" above its center cell), or (probably less economically justifiable) N by N or N by M convergence. Convergence might vary from wafer to wafer. In some cases there might be no convergence at all, or even divergence.

Especially considering that a wafer is a round disk (in contrast to the square chips into which it is usually sliced), the overall shape of any of the above systems might be a cylinder (rather than a square lattice) or a cone (rather than a pyramid). This will also be preferable whenever we can assume that the scene to be processed has a "center of interest" and that information becomes decreasingly pertinent as its distance from that center increases (a quite reasonable assumption for perception or modelling of masses of matter like storms or earthquakes).

MIMD Lattices, Trees and Networks of Stacked and Tiled Chips

Up to this point I have assumed each system has:
1) one controller, giving an SIMD computer where every processor executes the same instruction at each moment;
2) all processors with the same architecture (except for the special-purpose processor that collects information from adjacent processors);
3) stacks of whole wafers, where each wafer contains an array.

A number of interesting additional variations become possible if these assumptions are relaxed:

MIMD Pyramids and Lattices of SIMD Arrays

Each wafer might be given its own controller. This would allow for different instructions to be executed at each layer of either a pyramid or lattice system. Such a possibility seems highly desirable, since, effectively, it allows the whole system to be used as a 2-dimensional pipeline (actually, a "pipe-network"), with each layer executing the next instruction in the pipe. It is also quite useful for the closely related pipeline-like sequencing of image processing and perceptual transformations effected by a number of pattern recognition and scene description systems (e.g., Hanson and Riseman, 1974; Uhr, 1972, 1978; Uhr and Douglass, 1979; Schmitt, 1981).

The cost of the additional controllers might be compensated for by the simpler design and linkage of a controller that was on the same wafer as the processors it controlled, as well as the

much greater freedom in coding a greater variety of algorithms, and the higher percent of processors executing useful instructions.

Pyramids and Lattices of Different Types of Computers

Each wafer might have an array of (different numbers of) different types of computers. Here again there are several possibilities (see Uhr, 1983a, 1983b):

A) Moving "up" and "into" the pyramid, each wafer might have successively more powerful computers. This would now allow successively greater percentages of the wafer area to be filled with active procesor gates (see Uhr, 1982) that perform useful work. For example, when there was a 2 by 2 to 1 convergence of 4 computers the parent computer might be given 4 times as many gates.

B) Specialized layers could be built, as deemed appropriate. For example, a first layer of processors specialized to perform 8-bit arithmetic operations to smooth or enhance the raw input image might be placed directly where TV images were input to a perception system.

C) Limited reconfiguring might be used to allow the programmer to specify whether 1-bit, 8-bit, 16-bit or 32-bit processors were to be used (see Uhr, Thompson and Lackey, 1980). Other types of local reconfiguring (e.g., as in CLIP4 where either hexagonal or square array connections can be used, under program control) might be implemented. Global reconfiguring, between wafers or chips, might be added, as in PASM (Siegel, 1981), to handle needed remappings of data.

D) Within each wafer there might be several controllers, each responsible for a sub-region. This might be especially useful at the so-called "higher" or deeper levels of a perceptual system, where each sub-region of computers could now be applying the different set of operations suggested by what the previous SIMD operations had found. These might either be (possibly relatively more powerful) single computers, or local SIMD sub-regions on the larger MIMD wafer (reminiscent of Siegel's 1981 PASM, and Uhr, Thompson and Lackey's, 1980 Array/Net).

Trees and Networks of Wafers and Chips

A final set of possibilities (some of which might prove quite unattractive, or even virtually impossible, depending upon details of packaging technology) involve breaking away from the exact stacking of wafers, which to this point has been assumed to be as though into a nicely stacked loaf of sliced bread.

A) The wafers might be partially overlapped, giving 3-dimensional stacks of layers, each layer containing several wafers in a brick wall (but 3-dimensional) tiling-like pattern. For example (assuming square wafers for simplicity), a 2 by 2 of 4 wafers might have each wafer's "inner quarter" linked to the computers in a single wafer in an adjacent layer. Round or hex-

agonal wafers (probably of hexagonal arrays) would lend themselves especially well to such tilings. (The gaps between them would serve well for heat dissipation.) Now far larger systems become possible, without potential limits, since such tilings of wafers could expand indefinitely in any or all layers. Computers on the borders of wafers might link through the adjacent wafers in adjacent layers when links between neighbor wafers in the same layer were not feasible.

B) Any or all of the wafers might be diced into (square or hexagonal) chips, and chips used, as in A) above, to build larger structures. This would allow not only for massive tiling, but also for a variety of other tree and network configurations. This will become an increasingly attractive alternative, as packing densities increase to, and beyond, the point where each chip can contain a whole array or other powerful module.

C) With the whole system modularized (where desired) into chips, we can think in terms of more design flexibilities, including chips with different types of processors, specialized as deemed appropriate, and chips, or local sets of chips, with their own controllers.

Wafers or chips could be spaced to best handle heat dissipation and structural stability (since the spacers between layers would form an interior skeleton that would strengthen the mosaic). Thus we could begin to think of great walls, pyramids, and other appropriately sculpted 3-dimensional monoliths many wafers big in all three dimensions.

Summary Comments

Etchells, Grinberg and Nudd are developing what may well be the key to enormous increases in computer power. If they are able to transfer to production their already successful laboratory techniques for thermomigration of feedthroughs in silicon chips, and for microspring connections between two adjacent wafers, it will for the first time be possible to design very large multi-computer systems without the severe bottlenecks of pin fanout technology.

This paper briefly explores some of the architectures that this exciting new development may make possible. These include not only very large SIMD arrays and 3-dimensional lattices and pyramids, but also MIMD systems of pyramids, cones, lattices, cylinders, trees, tilings, and a variety of other network structures.

In particular, 3-dimensional tilings seem especially attractive. Such systems would tile either wafers or chips into layers of potentially any size. Then these layers would be stacked, to potentially any depth. Adjacent wafers or chips in the same layer could now be linked via adjacent layers, when desirable. Since tilings would not have to be dense, it would now be possible to construct virtually any network topology. Heat dissipation could be controlled nicely by suitable spacing between

layers and within layers. The structure of the total system would also be strengthened by the interspersed interior placement of the spacers at the borders of the wafers or chips, much as the structure of a wall is strengthened when its bricks or tiles are staggered and/or given rimmed edges.

Consider what might be achieved with the already feasible basic building-block module of a wafer (or chip) with a 32 by 32 array of 1024 computers (or a similar number of computers on a round or hexagonal wafer).

1) This might be realized in a single wafer or chip, or, as in the 3-D computer, on 10 or 20 stacked wafers.

2) Then a 128 by 128, or a 512 by 512 image processing-perception system could be built by tiling a 4 by 4, or a 16 by 16, array of such modules into a (first) layer.

3) Finally, 4 to 16 additional layers, each either the same size or successively smaller, could be stacked on top, to give a 3-D lattice or pyramid.

As VLSI packing grows denser, the individual computer might be given more power and/or more memory; and/or more computers might be placed on each module.

References

Batcher, K.E., Design of a massively parallel processor, IEEE Trans. Computers, 1980, 29, 836-840.

Duff, M. J. B., Review of the CLIP image processing system, Proc. AFIPS NCC, 1978, 1055-1060.

Etchells, R.D., Grinberg, J. and Nudd, G.R., The development of a novel three-dimensional microelectronic processor with ultra-high performance, Proc. Soc. Photo-Optical Instrumentation Engin., 1981.

Hanson, A. R. and Riseman, E. M., Pre-processing cones: a computational structure for scene analysis. COINS Tech. Rept. 74C-7, Univ. of Mass., 1974.

Kuck, D.J., The Structure of Computers and Computation: Vol. 1, New York: Wiley, 1978.

Kung, H.T., The structure of parallel algorithms. In: Advances in Computers, Vol. 19, M.C. Yovits (Ed.), 1980, 293-326.

Levine, M. D., A knowledge-based computer vision system. In: Computer Vision Systems, A. Hanson and E. Riseman (Eds.), New York: Academic Press, 1978, 335-352.

Patterson, D.A., Fehr, E.S., and Sequin, C.H., Design considerations for the VLSI processor of X-tree, Proc. Sixth Ann. Symp. on Computer Arch., 1979, 90-101.

Reddaway, S.F., DAP - a flexible number cruncher, Proc. 1978 LASL Workshop on Vector and Parallel Processors, Los Alamos, 1978, 233-234.

Rosenfeld, A., Quadtrees and pyramids for pattern recognition and image processing, Proc. Fifth Int. Conf. on Pattern Recognition, 1980, 802-807.

Schmitt, L., The ICON perception laboratory user manual and reference guide, Computer Sciences Dept. Tech. Rept. 421, 1981.

Siegel, H.J., PASM: a reconfigurable multimicrocomputer system for image processing. In: Languages and Architectures for Image Processing, M. J. B. Duff and S. Levialdi (Eds.), London: Academic Press, 1981.

Tanimoto, S. L., Pictorial feature distortion in a pyramid, Computer Graphics Image Proc., 1976, 5, 333-352.

Tanimoto, S. L. and Klinger, A. (Eds.), Structured Computer Vision: Machine Perception Through Hierarchical Computation Structures, New York: Academic Press, 1980.

Uhr, L., Layered "recognition cone" networks that preprocess, classify and describe, IEEE Trans. Computers, 1972, 21, 758-768.

Uhr, L., "Recognition cones" and some test results. In: Computer Vision Systems, A. Hanson and E. Riseman (Eds.) New York: Academic Press, 1978, 363-372.

Uhr. L., Comparing serial computers, arrays and networks using measures of "active resources," IEEE Trans. Computers, 1982, 30, 1022-1025.

Uhr, L., Algorithm-Structured Computer Arrays and Networks: Parallel Architectures for Perception and Modelling, New York: Academic Press, 1983, in press. (a)

Uhr, L., Pyramid Multi-Computer Structures, and Augmented Pyramids, In: Computing Structures for Image Processing, M. J. B. Duff, Ed., London: Academic Press, 1983, in press. (b)

Uhr, L. and Douglass, R., A parallel-serial recognition cone system for perception, Pattern Recognition, 1979, 11, 29-40.

Uhr, L., Lackey, J. and Thompson, M., A 2-layered SIMD/MIMD parallel pyramidal "array/net," Computer Sci. Dept. Tech. Rept. 409, Univ. of Wisconsin, 1980.